Algorithm 1 Stable Monotonic Chunkwise Attention Decoding

```
Input: encoder features $H = \{h_1, \cdots, h_U\}$, output index $i$, decoder hidden state $s_i$, output label $y_i$, endpoint $t_i$, sigmoid function $\sigma(\cdot)$, attention chunk width $w$
Initialize $s_0 = \vec{0},\ y_0 = \langle sos \rangle,\ t_0 = 1,\ i = 1$
while $y_{i-1} \neq \langle eos \rangle$ do
    for $j = t_{i-1}$ to $U$ do
        $e_{i,j} = g\frac{v_m^\top}{||v_m||}\tanh(W_m^s s_{i-1} + W_m^h h_j + b_m) + r$
        $p_{i,j} = \sigma(e_{i,j})$
        if $p_{i,j} \geq 0.5$ then
            for $k = j - w + 1$ to $j$ do
                $u_{i,k} = v_c^\top \tanh(W_c^s s_{i-1} + W_c^h h_k + b_c)$
            end for
            $c_i = \sum_{k=j-w+1}^{j} \frac{\exp(u_{i,k})}{\sum_{l=j-w+1}^{j} \exp(u_{i,l})} h_k$
            $t_i = j$
            break
        end if
    end for
    if $p_{i,j} < 0.5,\ \forall j \in \{t_{i-1}, \cdots, U\}$ then
        $c_i = \vec{0}, \quad t_i = t_{i-1}$
    end if
    $y_i \sim \text{Decoder}(s_{i-1}, y_{i-1}, c_i), \quad i = i + 1$
end while
```

Algorithm 2 Stable Monotonic Chunkwise Attention Training

```
Input: encoder features $H = \{h_1, \cdots, h_U\}$, output index $i$, decoder hidden state $s_i$, output label $y_i$, sigmoid function $\sigma(\cdot)$, attention chunk width $w$, Gaussian noise $\epsilon$
$s_0 = \vec{0},\ y_0 = \langle sos \rangle,\ \alpha_{0,0} = 1,\ \alpha_{0,k} = 0(k \neq 0),\ i = 1$
while $y_{i-1} \neq \langle eos \rangle$ do
    for $j = 1$ to $U$ do
        $e_{i,j} = g\frac{v_m^\top}{||v_m||}\tanh(W_m^s s_{i-1} + W_m^h h_j + b_m) + r$
        $p_{i,j} = \sigma(\text{Energy}(s_{i-1}, h_j) + \epsilon)$
        $\alpha_{i,j} = p_{i,j} \prod_{k=1}^{j-1}(1 - p_{i,k})$
    end for
    for $j = 1$ to $U$ do
        $u_{i,j} = v_c^\top \tanh(W_c^s s_{i-1} + W_c^h h_j + b_c)$
        $\beta_{i,j} = \sum_{k=j}^{j+w-1} \frac{\alpha_{i,k}\exp(u_{i,j})}{\sum_{l=k-w+1}^{k} \exp(u_{i,l})}$
    end for
    $c_i = \sum_{j=1}^{U} \beta_{i,j} h_j$
    $y_i \sim \text{Decoder}(s_{i-1}, y_{i-1}, c_i), \quad i = i + 1$
end while
```

Algorithm 3 Monotonic Truncated Attention Decoding

```
Input: encoder features $H = \{h_1, \cdots, h_U\}$, output index $i$, decoder hidden state $s_i$, output label $y_i$, endpoint $t_i$, sigmoid function $\sigma(\cdot)$, attention chunk width $w$
Initialize $s_0 = \vec{0},\ y_0 = \langle sos \rangle,\ t_0 = 1,\ i = 1$
while $y_{i-1} \neq \langle eos \rangle$ do
    for $j = 0$ to $U$ do
        $e_{i,j} = g\frac{v_m^\top}{||v_m||}\tanh(W_m^s s_{i-1} + W_m^h h_j + b_m) + r$
        $p_{i,j} = \sigma(e_{i,j})$
        $\alpha_{i,j} = p_{i,j} \prod_{k=1}^{j-1}(1 - p_{i,k})$
        if $p_{i,j} \geq 0.5$ then
            $c_i = \sum_{k=1}^{j} \alpha_{i,k} h_k$
            $t_i = j$
            break
        end if
    end for
    if $p_{i,j} < 0.5,\ \forall j \in \{t_{i-1}, \cdots, U\}$ then
        $c_i = \vec{0}, \quad t_i = t_{i-1}$
    end if
    $y_i \sim \mathrm{Decoder}(s_{i-1}, y_{i-1}, c_i), \quad i = i + 1$
end while
```

Algorithm 4 Monotonic Truncated Attention Training

```
Input: encoder features $H = \{h_1, \cdots, h_U\}$, output index $i$, decoder hidden state $s_i$, output label $y_i$, sigmoid function $\sigma(\cdot)$, attention chunk width $w$, Gaussian noise $\epsilon$
$s_0 = \vec{0},\ y_0 = \langle sos \rangle,\ \alpha_{0,0} = 1,\ \alpha_{0,k} = 0(k \neq 0),\ i = 1$
while $y_{i-1} \neq \langle eos \rangle$ do
    for $j = 1$ to $U$ do
        $e_{i,j} = g\frac{v_m^\top}{||v_m||}\tanh(W_m^s s_{i-1} + W_m^h h_j + b_m) + r$
        $p_{i,j} = \sigma(\mathrm{Energy}(s_{i-1}, h_j) + \epsilon)$
        $\alpha_{i,j} = p_{i,j} \prod_{k=1}^{j-1}(1 - p_{i,k})$
    end for
    $c_i = \sum_{j=1}^{U} \alpha_{i,j} h_j$
    $y_i \sim \mathrm{Decoder}(s_{i-1}, y_{i-1}, c_i), \quad i = i + 1$
end while
```